MTCNN-UGAN: A Self-Attention Enhanced Face Replacement Pipeline for Film and Television Video Frames

Yumeng Wu, Xiao Wang*
School of Art, Zhejiang Yuexiu University, Zhejiang, 312000, China
E-mail: ssshayuya@163.com

Keywords: MTCNN, face replacement, adversarial generative networks, self-attention, digital de-jittering

Received: April 16, 2025

At present, face replacement technology in film and television videos faces problems such as low accuracy and high resource consumption. Research proposes an automated face replacement technology that integrates improved multi task cascaded convolutional neural networks (MTCNN) and generative adversarial networks (GAN). In the face detection stage, MTCNN is used, and a median filter preprocessing and depthwise separable convolution model are introduced. In the face replacement stage, a U-Net based generative adversarial network (UGAN) is constructed, whose generator consists of an encoder and a decoder, and is embedded with a dual skip connection residual module. The discriminator adopts a self attention mechanism and a video stabilization module. In the experiment, WIDER FACE and Celeb Faces Attributes Dataset (CelebA) were used for face detection tasks. The face replacement task used a high-resolution Celebrity Mask High Quality (CelebAMask HQ) dataset and a Deepfake Model Attribution Dataset (FDM). Meanwhile, the study introduced FaceSwap technology and attribute preserving generative adversarial network (AP-GAN) as comparative baselines. In face detection experiments, the research model performed best in terms of accuracy as well as training loss in different face detection scenes. For example, the accuracy of the research model in complex scenes was 93.25%, and the training loss was 0.221. In the face replacement experiment, the model replaces faces in four image sets. Its color as well as face contour structure was well preserved and face replacement was more natural. In the similarity index comparison, the research model performed the highest face replacement similarity index at different frame numbers with an average value of 0.994. The research model also performed the best in the face replacement imaging peak signal-to-noise ratio test with an average value of 35.65. Finally, in the face replacement composite test, the research model performed the best in both structural similarity and state error. In conclusion, the technique has good application results. This study can provide technical support for the improvement of face replacement technology as well as face characterization.

Povzetek: vtorja združita izboljšano zaznavanje obrazov (predhodno čiščenje šuma, lažji konvolucijski sloji in večločnostni prikaz), generator v slogu U-Net z dvojnimi povezavami, pametnejši diskriminator s samousmerjanjem pozornosti ter modul za stabilizacijo videa (odprava tresenja). Rezultat so naravni prenosi videza, visoka podobnost z izvirnikom in tekoče predvajanje tudi pri zahtevnih prizorih.

1 Introduction

To create a very realistic look, face-swapping technology is a type of picture and video processing tool that can swap out one person's face for another [1]. The technology originated from early image editing. Computer vision technology has advanced quickly in recent years and is now widely employed in advertising production, special effects in movies and television shows, and other sectors [2]. At present, with the continuous development of deep learning (DL), artificial intelligence (AI) and other technologies, the related face-swapping technology ushers in rapid development and attracts the attention of a large number of scholars [3]. To improve the application of face exchange technology, Rao et al. proposed an approach based on convolutional god coding and decoding network. In it, face marker point detection and alignment were combined with clustering and computer vision techniques, and a large amount of face data was trained to construct a face model. Face clusters were generated by clustering to optimize the face exchange effect. It was shown through experiments that this technique had good application effect, but it was poorly applied in low-end devices [4]. Omar K et al. proposed a DL bagging based integrated classifier for the deep forgery video detection problem, which employed convolutional and self-attentive networks as the basic learners. The model vertically stacked the deep convolutional and self-attention layers and extracted the local features of the face from the video and trained by learning to achieve face replacement. Finally, the study trained the technique on a public dataset. The results indicated that the technique had good video processing capability and high training accuracy [5]. Abdelminaan et al. studied the problem of detecting deep-fake videos. They developed a

web application that detected the authenticity of video input and protected public figures and politicians from false videos. The research adopted the method of combining machine learning and DL to analyze the data set containing deep forgery and real video. This method could effectively distinguish between real and forged content, such as face replacement or voice replacement. The results could be used in courts and police stations to reduce the risk of crime and fraud, while improving the detection efficiency and providing guarantee for the credibility of network information [6].

AI has made incredible strides in image and video processing thanks to the quick development of DL technology. Among them, video face replacement technology, as an emerging image and video editing technology, utilizes DL algorithms and neural networks to achieve highly realistic face replacement effects. Tsai C S et al. put up a useful framework for enhancing angle transformation and face replacement in order to address the issue of face replacement distortion. The framework contained a transformation framework based on generative adversarial networks, and generated multi-angle transformed images by combining the predicted face points to realize face recognition. The method's ability to preserve high-quality photos and prevent image distortion during face replacement with image angle changes was demonstrated in experimental tests [7]. Melnik et al. conducted a comprehensive review of DL methods for generating and editing faces in StyleGAN. The study analyzed the evolution of StyleGAN from PGGAN to StyleGAN3 and discussed key issues such as training indicators, potential representations, GAN inversion, and face image editing. The research also involved cross-domain face styling and restoration. This research served as an entry point into the field of face generation. It helped beginners quickly understand related technologies and promoted the development of face generation and editing technology. Additionally, it provided an important reference for subsequent research. Liao X et al. investigated the problem of compressed depth fake video detection in social networks and proposed a detection framework that considered facial muscle movement to realize the detection of face-swapped videos. The framework achieved this by localizing faces from consecutive frames, extracting marker points, and modeling sensory regions and face regions. Experimental results demonstrated that the method outperformed existing methods in detecting compressed depth pseudo-video [9]. The problem of deep face-swapping technique detection is studied by Akhtar Z et al. In this, a comprehensive review of existing images, videos, and Deepfake databases was conducted to propose a deep face-swapping detection framework. It adopted a new generation of deep feature point recognition technology, which was trained by a large amount of face data to realize false video detection. The experimental results indicated that the technique had a good detection accuracy, but the shortcomings of the technique were the lack of a unified detection standard and the poor applicability to the detection of new type of face data [10]. The comparison of relevant literature research is shown in Table 1.

Table 1: Literature review research

Researcher	Research contents	Comparison of research results with FaceSwap, DeepFaceLab, AP-GAN, etc	Compared with FaceSwap, DeepFaceLab, AP-GAN, etc
Rao I S S et al. [4]	Facial swapping technology based on convolutional neural networks, combined with clustering and computer vision techniques	The application effect is good, but it performs poorly on low-end devices	Compared with FaceSwap, it optimizes face alignment but has a higher computational complexity
Omar K et al. [5]	Deep fake video detection based on deep learning bagging ensemble classifier	Strong video processing ability and high training accuracy	Compared to DeepFaceLab, it focuses more on extracting local video features, but the model is complex
Abdelminaam d et al. [6]	It studies face video forgery, and detects it combined with related voice and portrait images	High accuracy and good efficiency, but high requirements for hyperparameter adjustment	Compared with traditional AP-GAN, it optimizes computational efficiency but requires higher demands on the dataset
Tsai C S et al. [7]	An effective framework for improving angle transformation and face replacement based on generative adversarial networks	Maintain high quality during image angle transformation to avoid distortion	Compared to traditional FaceSwap, it solves the problem of image distortion, but the training difficulty is higher
Melnika et al. [8]	Review and analyze the current technologies related to face generation and editing, and analyze the effects of different technologies	The detection effect is superior to similar models	Compared with the benchmark DeepFaceLab, the detection accuracy is higher,

			but the model dependency is strong
Liao X et al. [9]	A compressed depth pseudo video detection framework considering facial muscle movement	The effect of detecting compressed depth pseudo video is better than existing methods	Compared to the benchmark AP-GAN, it is more suitable for compressed video detection, but its generalization ability is limited
Akhtar Z et al. [10]	Deep face swapping detection framework, utilizing next-generation deep feature point recognition technology	High detection accuracy, but lack of unified standards, poor applicability to new data	Compared to FaceSwap and DeepFaceLab, it has higher detection accuracy, but its applicability is limited

According to the above research, with the development of DL as well as AI and other technologies in recent years, video face-swapping technology has seen rapid development. It has been used in video creation, face data detection and other fields. However, according to the above study, the current video face replacement technology still faces many shortcomings. For example, the low similarity of face replacement, the high consumption of face replacement technology resources, and the poor adaptation of complex scene technology all limit the development of the technology. Current research problems include insufficient adaptability to complex scenes. For example, there is a significant decrease in detection accuracy under occlusion and lighting changes. The consumption of computing resources is too high. For example, traditional multi task cascaded convolutional neural networks (MTCNN) detection models have a large number of parameters, which makes feature pyramid calculation inefficient and difficult to meet the real-time processing requirements of film and television videos. Therefore, in order to solve the face replacement consumption as well as accuracy problems, the study proposes a video face replacement technology based on improved MTCNN. There are two innovations in the research. One is to adopt the improved MTCNN algorithm, which improves face detection accuracy and efficiency through optimization such as median filtering and depth separation convolution. Second, the research constructs face replacement model based on adversarial generative network. It adopts the self-attention mechanism (SAM) and digital de-jittering method to improve the face replacement accuracy and video smoothness. This research can provide technical support for the improvement of video face replacement technology.

2 Methods and materials

2.1 Face detection model for film and television videos based on improved MTCNN algorithm

With the rise of AI as well as video creation, film and television works face replacement technology is more and more sought after. However, the traditional technology is inefficient and the replacement color deviation is large, which cannot meet the requirements of film and television videos creation. In this regard, the study proposes a face replacement technology that combines MTCNN algorithm and adversarial generative network. The flowchart of the whole technology is shown in Figure 1.

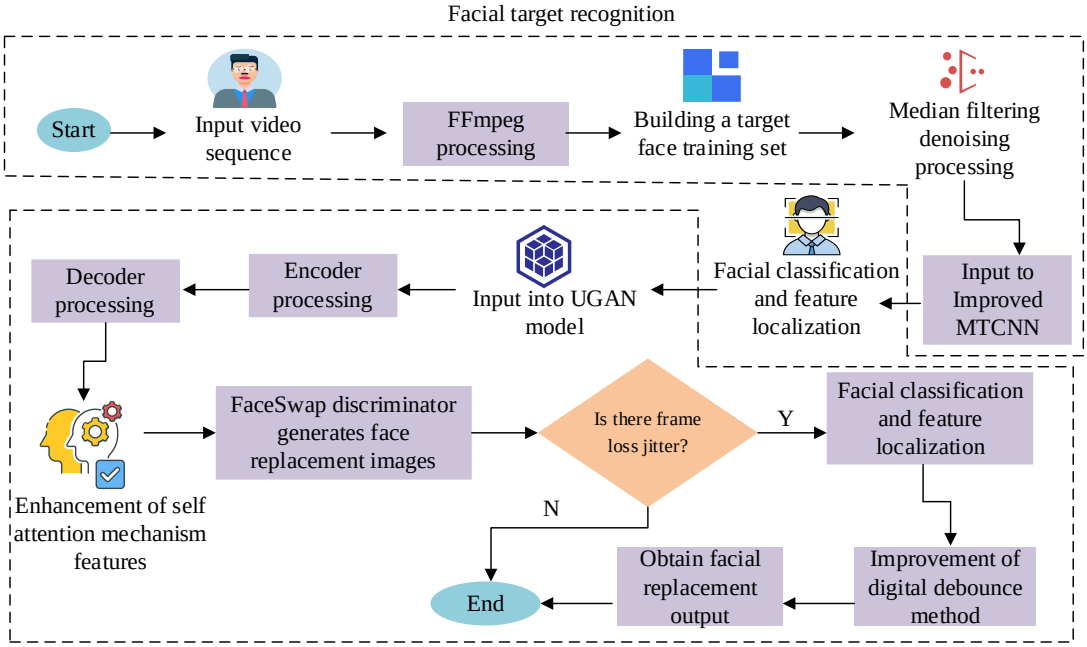


Figure 1: Process of automatic face replacement technology in film and television videos

In Figure 1, the technique has two parts. The first part is the face feature detection part and the second part is the face replacement based on adversarial generative networks. The face detection part of the study uses MTCNN algorithm as face detection. It has high face detection accuracy, suitable for all types of face feature recognition, and excellent performance in the face detection field. Figure 2 depicts the MTCNN's three-stage structure.

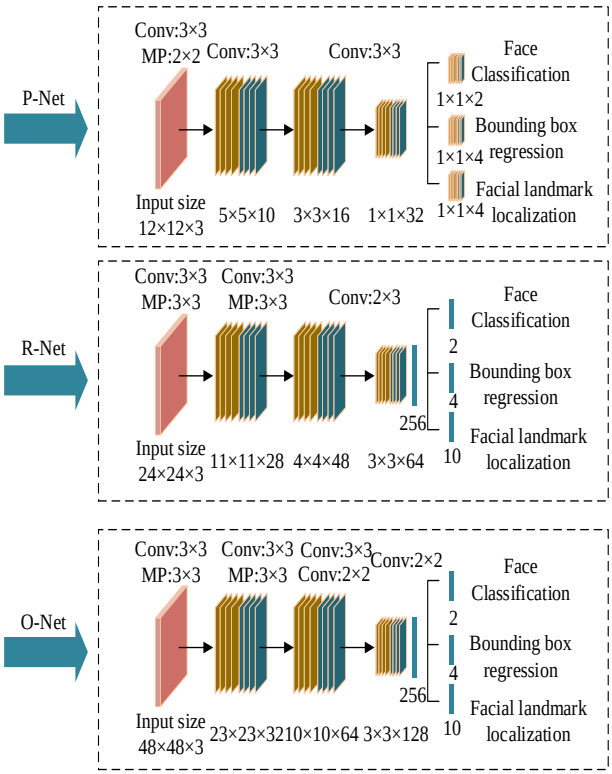


Figure 2: Three stage structure of MTCNN

According to the MTCNN, the network uses P-net network to recognize the face region in the first stage of face recognition. It mainly achieves face feature classification as well as feature localization through multilayer convolutional processing [11]. The R-net network is used in the second stage to detect faces. It eliminates non-face candidate frames through non-maximum suppression (NMS) and other processing methods, and strengthens the analysis of face features, localization data, etc [12]. O-Net network is used in the third stage. This process refines the bounding box detection region and adds new convolutional layer processing. In the process of face classification, the image will be recognized whether it is a face or not. Equation (1) illustrates how a cross-entropy loss function represents the process [13].

$$L_i^{det} = -\left(y_i^{det}\log\left(p_i\right)+\left(1-y_i^{det}\right)\left(1-\log\left(p_i\right)\right)\right) \quad (1)$$

In Equation (1), p_i represents the network recognition as face probability output. $y_i^{det} \in \{0,1\}$ is the sample detection label and the output is 1 or 0, 0 is non-face and 1 is face. Whereas, in face bounding box processing, Euclidean loss is used to reflect the candidate library detection, as shown in Equation (2).

$$L_i^{box} = \hat{y}_i^{box} - y_i^{box}{}_2^2 \quad (2)$$

In Equation (2), y_i^{box} represents the bounding box of the face in detection. $\hat{y}_i^{box}$ represents the coordinates of the predicted corrected bounding box. Next, the minimum Euclidean loss is used in face localization to reflect the location of key features of the recognized face, as shown in Equation (3).

$$L_i^{landmark} = \hat{y}_i^{landmark} - y_i^{landmark}{}_2^2 \quad (3)$$

In Equation (3), $y_i^{landmark}$ is the real human face feature key point coordinates. $\hat{y}_i^{landmark}$ represents the

predicted face feature keypoint coordinates. Then the MTCNN loss optimization is obtained based on the above analysis, as shown in Equation (4) [14].

$$min\sum_{i=1}^{N}\sum_{j\in\{det,box,landmark\}}\alpha_j\beta_i^j L_i^j \quad (4)$$

In Equation (4), a_j represents the weight set without loss, which reflects the degree of influence of the loss function. Its weight set is 1. The face classification weight is 0.3. The bounding box regression loss weight is also 0.3. The keypoint localization loss weight is 0.4. N represents the total number of training face samples $\beta_i^j \in \{0,1\}$ represents the sample type indicator. 1 is face and 0 is non-face. Although MTCNN has excellent recognition effect on face detection compared to the traditional target detection model, the traditional MTCNN still faces the problems of high resource consumption, low detection efficiency, and low processing effect on noisy images. For example, before constructing the face feature map, some of the images captured by the camera system contain noise such as Gaussian and pretzel, which affects the network detection effect. Therefore, the input samples are pre denoised and then fed into the MTCNN model for training [15]. The median filtering process is shown in Equation (5).

$$g(x,y)=median\{f(x-m,y-n)\,|\,(m,n)\in W\} \quad (5)$$

In Equation (5), $f(x-m,y-n)$ denotes the filtered pixel position with offset. $g(x,y)$ denotes the original map pixel position. W denotes the filter window. $median$ is the median value of the filter element. m and n are the window offset parameters. In MTCNN, network will extract the features of each layer of the image pyramid to obtain multi-scale image features, but the process is computationally intensive. To solve the problem, the study considers image pyramid and feature map pyramid. The former predicts only the high-level features in multilayer feature extraction, which improves the efficiency but decreases the detection accuracy [16]. Unlike the image pyramid, the latter considers different layer feature prediction relationships in the multiscale feature output stage. This reduces the amount of computation while ensuring training accuracy. Therefore, the feature pyramid is integrated prior to MTCNN inference in order to construct an image pyramid of the input image and generate a multi-scale image sequence. Finally, the study also introduces depth-separated convolution to improve the MTCNN convolutional operation parameter problem and enhance the network detection efficiency. Moreover, depthwise depth-separated is integrated into the convolutional layers of P-Net, R-Net, and O-Net to reduce computational and parameter complexity. For example, in R-Net/O-Net, depth-separated is applied to all convolutional layers. For example, in R-Net, the input features are first deeply convolved and then combined across channels via 1×1 convolution to output the bounding box and keypoint coordinates. The activation function adopts LeakyReLU instead of ReLU to enhance the retention of negative value information. The MTCNN structure is shown in Figure 3.

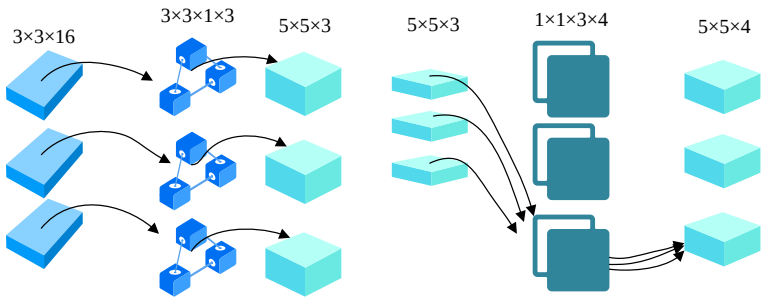


Figure 3: Depth separable convolution

Figure 3 shows the depth separable convolution process (CP). It resizes the input 7×7×3 feature map to 5×5×4 size by convolving with channel number 1. Compared to the standard CP, it preserves the number of image feature channels and guarantees the quality of the graph while reducing the height and width of the feature graph. In this, the depth convolution is shown in Equation (6).

$$Y_{dw}^{(c)}(x,y)=\sum_{i=k}^{k}\sum_{j=-k}^{k}X^{(c)}(x+i,y+j)\cdot K_{dw}^{(c)}(i,j)+b_{dw}^{(c)}(\quad (6)$$

In Equation (6), $X^{(c)}(x+i,y+j)$ denotes the value of the output tensor C th channel position $(x+i,y+j)$. $K_{dw}^{(c)}(i,j)$ denotes the (i,j) th position weight of the deep convolution at C channel. $b_{dw}^{(c)}$ denotes the C th channel bias of the depth convolution. $Y_{dw}^{(c)}(x,y)$ denotes the output at channel position (x,y). Next, the point-by-point convolution operation is shown in Equation (7).

$$Y_{pw}^{(c)}(x,y)=\sum_{c=1}^{C_{in}}Y_{dw}^{(c)}(x,y)\cdot K_{pw}^{(c,c')}+b_{pw}^{(c')} \quad (7)$$

In Equation (7), $Y_{dw}^{(c)}(x,y)$ denotes the value of the deep convolutional C th channel at position (x,y). $K_{pw}^{(c,c)}$ denotes the weight of the point-by-point

convolution kernel from the input channel C to the output channel C'. $b_{pw}^{(c')}$ denotes the point-by-point convolutional C'th output bias. $Y_{pw}^{(c)}(x,y)$ denotes the value of the final output C'th channel at position (x,y). Finally, to improve the training accuracy of MTCNN face replacement classification task, SoftMax loss function is used to replace the cross loss function. Finally, to improve the training accuracy of the MTCNN network for face replacement classification tasks, the SoftMax loss function is introduced, based on the original cross-loss function. The squared-difference loss function is used to evaluate the performance of face-frame detection predictions. The SoftMax loss function is shown in Equation (8) [17].

$$SoftMax(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_i}} \quad (8)$$

In Equation (8), x_i denotes sample i and y_i denotes labeled values. The squared deviation loss function is shown in Equation (9).

$$Loss == \frac{1}{N'} \sum i = 1^{N'} \beta_i \cdot \left(\| y\_pred^{(i)} - y\_gt^{(i)} \|_2^2 \right) (9)$$

In equation (9), $y\_pred^{(i)}$ represents the predicted value, $y\_gt^{(i)}$ represents the true value, β_i represents the sample type indicator, and N' represents the sample size.

2.2 Adversarial generative network-based automatic face replacement model for film and television videos

After completing the recognition of face images, the next step is to perform automatic video face replacement modeling. This part uses adversarial generative networks as face replacement technology. Currently, adversarial generative networks have become a representative of the field of image synthesis and transformation. Compared with the traditional DeepFakes class of face replacement technology, its face replacement accuracy is high and can be automatically replaced for the input source. Therefore, the study proposes a novel automatic face switching generation framework for self-coding networks (U-Net) on the framework of deep convolutional adversarial networks (DCGAN), which is referred to as the UGAN model. The UGAN model mainly contains two parts: generator as well as discriminator. Figure 4 depicts the particular structure.

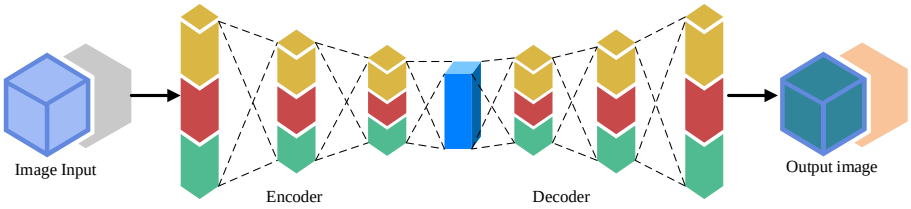


Figure 4: UGAN model generator structure

According to the structure of Fig. 4, there are two main components, Encoder and Decoder. In order for the generative model to output a specified realistic image according to the target, the feature information within the source image is extracted by the Encoder, including the background, face contours, etc. Moreover, this information is converted into a latent word feature vector. Then the Encoder processing vectors are input to the Decoder. It is mainly responsible for reconstructing the feature vectors, preserving more details of the target and outputting the reconstructed features. The Encoder in the generator first downsamples the input size of 256×256 source image to extract multiple face feature attributes. The UGAN Architecture with layer names, filter sizes, and skip connection types is in Table 2.

Table 2: UGAN Architecture with layer names, filter sizes, and skip connection types

Network	Layer Name	Enter size	Filter size/quantity	Step length	Fill in	Activation function
P-Net	Input	-	3 channels	-	-	-
Conv1	Any size (≥ 12 × 12)	10, 3×3	1	Valid	PReLU	-
MaxPool1	-	2×2	2	Same	-	-
Conv2	-	16, 3×3	1	Valid	PReLU	-
Conv3	-	32, 3×3	1	Valid	PReLU	-
Output (Cls/Reg)	-	2/4, 1×1	1	Valid	Softmax/Linear	Dual branch output
R-Net	Input	24×24×3	-	-	-	-
Conv1	24×24×3	28, 3×3	1	Valid	PReLU	-
MaxPool1	-	3×3	2	Same	-	-
Conv2	-	48, 3×3	1	Valid	PReLU	-

MaxPool2	-	3×3	2	Valid	-	-
Conv3	-	64, 3×3	1	Valid	PReLU	-
FC1	576 dimensions	Unit 128	-	-	PReLU	Fully connected layer
Output (Cls/Reg)	-	2D/4D	-	-	Softmax/Linear	Dual branch output
O-Net	Input	48×48×3	-	-	-	-
Conv1	48×48×3	32, 3×3	1	Valid	PReLU	-
MaxPool1	-	3×3	2	Same	-	-
Conv2	-	64, 3×3	1	Valid	PReLU	-
MaxPool2	-	3×3	2	Valid	-	-
Conv3	-	64, 3×3	1	Valid	PReLU	-
MaxPool3	-	2×2	2	Same	-	-
Conv4	-	12, 3×3	1	Valid	PReLU	-
FC1	1152 dimensions	256 units	-	-	PReLU	Fully connected layer
Output (Cls/Reg/Landmark)	-	2D/4D/1D	-	-	Softmax/Linear/Linear	Three task branches

Whereas in Decoder the up-sampling is done in the form of pixel CP to adjust the height and width $H \times W$ image to r times high resolution $rH \times rW$ image. Among them, r denotes the sampling factor times. Pixel rearrangement by Decoder is shown in Equation (10).

$$TS(T)_{x,y,c} = T_{\lfloor y/r \rfloor, \lfloor x,r \rfloor c \square r \square mod(y,r)+c \square mod(x,r)} \quad (10)$$

In Equation (10), $\mathrm{mod}(x, y)$ is the sampling coordinate position. An improved residual network model is added to the structure of the network when processing image data, especially when training numerous face data, in order to reduce the network's computational load. It makes use of a double jump connection topology, which can successfully address the gradient vanishing problem in data processing. The residual function expression is shown in Equation (11) [18].

$$F(x) = Conv_{3\times3}(Leaky\,\mathrm{Re}\,LU(Conv_{3\times3}(x))) \quad (11)$$

In Equation (11), x is the input information. $Conv$ is the convolutional processing. $LeakyReLU$ is the activation function. In this case, the input information x is processed twice using 3×3 convolutional processing to strengthen the feature reuse capability. Then, activation processing is performed by LeakyReLU activation. Next, the output information x is summed with the residual function for residual linkage. The study selects LeakyReLU activation for processing due to its high computational efficiency and suitability for real-time film and video processing requirements, despite its slightly slower convergence compared to GELU/Wish. LeakyReLU ensures stable convergence during training by avoiding neuronal death. The residual module is embedded in the encoder-decoder structure of the U-Net generator. This structure achieves cross-layer feature fusion through dual skip connections. It solves the problem of gradient vanishing and improves the ability to preserve facial details. At the same time, the robustness of forgery detection is enhanced by weighting key features in collaboration with the SAM in the discriminator. The residual module is shown in equation (12) [19].

$$\mathrm{ResBlock}(x) = x \oplus \mathrm{Conv}_{3\times3}(\mathrm{LeakyReLU}(\mathrm{Conv}_{3\times3}(x))) \quad (12)$$

Next, after completing the image feature extraction work through the generator, the study uses the network face replacement framework (FaceSwap) discriminator for face replacement processing. The framework mainly consists of a SAM layer as well as a convolutional layer. The structure of the discriminator framework is shown in Figure 5 [20].

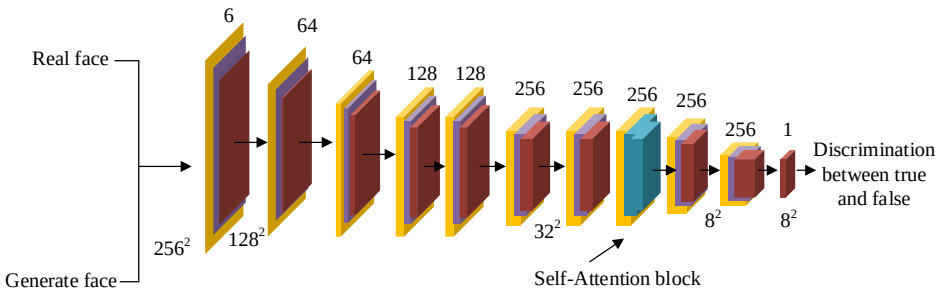


Figure 5: Discriminator framework structure

In Figure 5, in the discrimination framework, through iterative processing, the discriminator will discriminate the similarity between the generated face as well as the real face to ensure that the network generated image is closer to the real target face. Within the face replacement framework, a special layer of SAM is added. Its function is to weigh features based on their relationships in order to capture long-term dependencies and strengthen attention to key features. The mathematical expression of self attention mechanism is shown in equation (13).

$$\text{Attention}(Q,K,V)=\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (13)$$

In Equation (13), Q, K, and V denote the query vector, key vector, and value vector, respectively. d_k denotes the dimension of the key vector. d_k is the sequence length. After face replacement processing is completed by FaceSwap framework, the face replacement process still faces the video replacement jitter problem. In other words, during the face replacement session, the network performs replacement processing on multiple faces. This results in some video frames not being continuous, leading to video face loss and frame jitter. In this regard, the research introduces a digital de-jittering method. Its flow is shown in Figure 6.

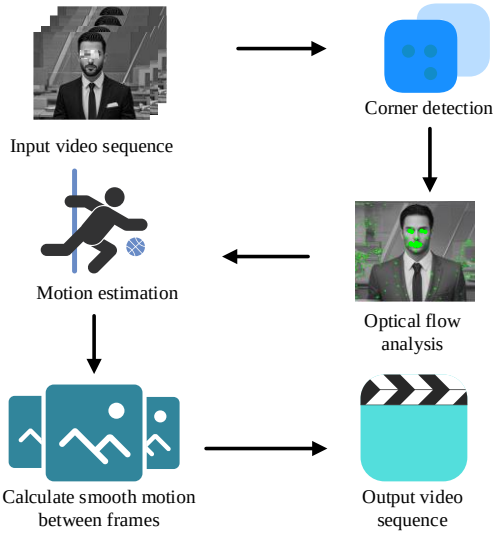


Figure 6: Process of face replacement video debounce

In Figure 6, this de-jittering process is realized through corner point detection, optical flow analysis, and motion estimation to achieve de-jittering. Among them, the process is realized by using Shi-Tomasi, a corner detection tool within OpenCV, which detects the gray level change anomalies in the image frame as corner points through a fixed window. Then optical flow analysis is performed for tracking the feature points in the next frame of the image. Lucas-Kanade is used as the analyzing method in the study, although it is not able to track all the moving points in some scenes. Therefore, the marker code mechanism is used to improve the problem. When the next frame tracking marker position is determined the status is marked as 1, otherwise 0 will re-update the tracking points. After completing the above analysis for motion estimation, the study uses random sample consistency (RSC) algorithm to estimate the variation between frames. The process uses Euclidean transform to reflect the superposition relationship of the features of each frame as shown in Equation (14).

$$S_E=\begin{pmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \\ 0 & 0 & 1 \end{pmatrix}\begin{bmatrix} X' \\ Y' \\ 1 \end{bmatrix}=T_E\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \quad (14)$$

In Equation (14), a_{11} denotes the translation factor.

a_{12} denotes the left-right rotation factor. a_{21} is the up-down rotation factor. a_{22} denotes the scaling factor. t_x and t_y are both transformation parameters. Next, the error between the front and back change frames is calculated for judging the estimated position. The sum of the errors between the front and back change frames is shown in Equation (15).

$$SAD = \sum_{i=1}^{N}\sum_{j=1}^{N}\left|c_{i,j} - r_{i,j}\right| \quad (15)$$

In Equation (15), $r_{i,j}$ represents the previous frame feature point location. $c_{i,j}$ represents the next frame feature point location. Finally, the study uses Smooth Filter to smooth the filter processing curve, which makes the video frame smoother. With the above processing, the automatic replacement of faces and the face replacement jitter problem are accomplished. The technical analysis has been completed through the above research, in which the improved MTCNN (Deep Separation Convolution+Feature Pyramid) and UGAN (Lightweight U-Net Generator) are synergistically optimized and meet the real-time running requirement of 30+FPS.

3 Results

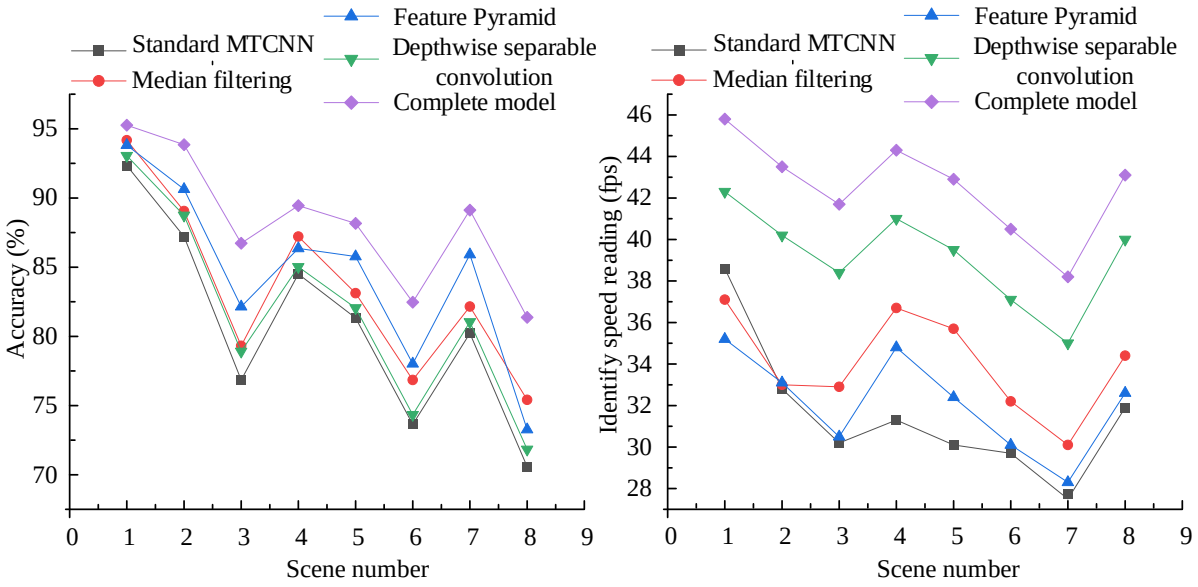
3.1 Face detection experiment based on improved MTCNN

Next, in order to test the proposed technology of the study, the corresponding face detection experiments and face replacement experiments will be carried out. Among them, the experimental system adopts WIDOWS 11 system, the processor adopts AMD Ryzen R5600, the graphics card is Nvidia RT3060, and the memory is 32GB DDR4 3200MHz. The experimental platform is Pyeharm 2021.1.1. The MTCNN face detection training time in the experimental training is 2.5 hours. The face replacement time is 28 hours, and an additional 3 hours are needed for debouncing. The parameter settings of the improved MTCNN are shown in Table 3.

Table 3: Model initial parameters

Parameter indicator type	Numerical value
Batch-size	16
Learning-rate	0.001
Weight	0.0005
Factor	0.7
Minsize	15

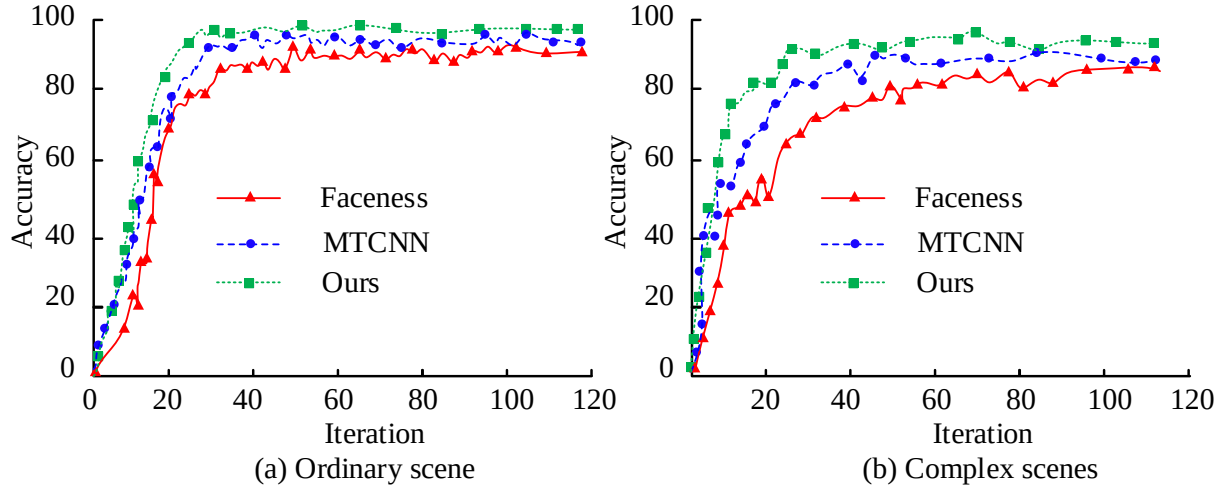
Table 3 shows the parameter settings for training the model, where Batch size is set to 16. A smaller batch size is selected to balance GPU memory limitations and training stability. This ensures that memory does not overflow and that training remains stable. To ensure stable training, Adam optimizer uses a standard learning rate of 0.001. The L2 regularization weight is set to 0.0005 to suppress overfitting. The image pyramid scaling factor is set to 0.7 to optimize the efficiency of multi-scale face detection. Minsize is set to 15, with a minimum facial pixel size of 15, to filter out noise and dryness. To effectively detect facial contours and provide a basis for analyzing information for subsequent face replacement, the professional face data WIDER FACE and the CelebFaces Attributes (CelebA) dataset are used to train the face detection model. The CelebA dataset mainly consists of a celebrity facial attribute dataset (with 40 attribute annotations such as glasses and hats), which includes 202599 facial attribute datasets and various types of facial image sets. The WIDER FACE dataset covers natural facial images of complex scenes (occlusion, lighting changes, blurring), with a total of 32203 images. First, the normal scene and complex scene (occlusion, dim light) of WIDER FACE dataset are selected for accuracy test. First, an ablation experiment is conducted. 8 types of image scenes are selected for analysis, including ordinary scenes, complex occlusion, small-scale faces, low lighting, large angle deflection, high dynamic blur, crowded scenes with multiple people, and strong noise interference, corresponding to image sequence numbers 1 to 8. The results of the ablation experiment are shown in Figure 7.



(a) Identification accuracy (b) Recognition rate

Figure 7: Results of ablation experiment

Figure 7(a) illustrates the recognition accuracy of the MTCNN model across various modules. According to the test results, standard MTCNN performs the worst in different scenarios, particularly in scenarios involving small faces and high dynamic models, achieving an accuracy below 80%. The best performing complete model has a performance rate of 87.6% and 83.6% in small-sized face and high dynamic model scenarios, respectively. Figure 7 (b) shows the comparison results of recognition rates. According to the test results, the complete model's average recognition rate is 42.5 fps. This is significantly better than the standard MTCNN's rate of 32.8 fps and better than a single MTCNN combination. In summary, the results indicate that adding modules such as median filtering and feature pyramids can significantly improve the model's application performance. Moreover, Faceness-Net (Faceness) and standard MTCNN are introduced as test benchmarks. The results are shown in Figure 8.



(a) Ordinary scene (b) Complex scenes

Figure 8: Accuracy of face detection in different scenarios

In Figure 8(a) for common scene detection, the fastest convergence of the research model compared to similar techniques is 98.35% accuracy at convergence of 30 iterations. Whereas MTCNN and Faceness converge with an accuracy of 92.31% and 82.65% respectively. In the complex scene detection in Figure 8(b), there is a significant fluctuation in Faceness face detection with the lowest overall accuracy of 81.25%. In comparison, the complex scene research model performs the best. The accuracy at convergence is 93.25% compared to 86.25% for MTCNN. Next, the same scene is compared to the training loss of different techniques, as shown in Figure 8.

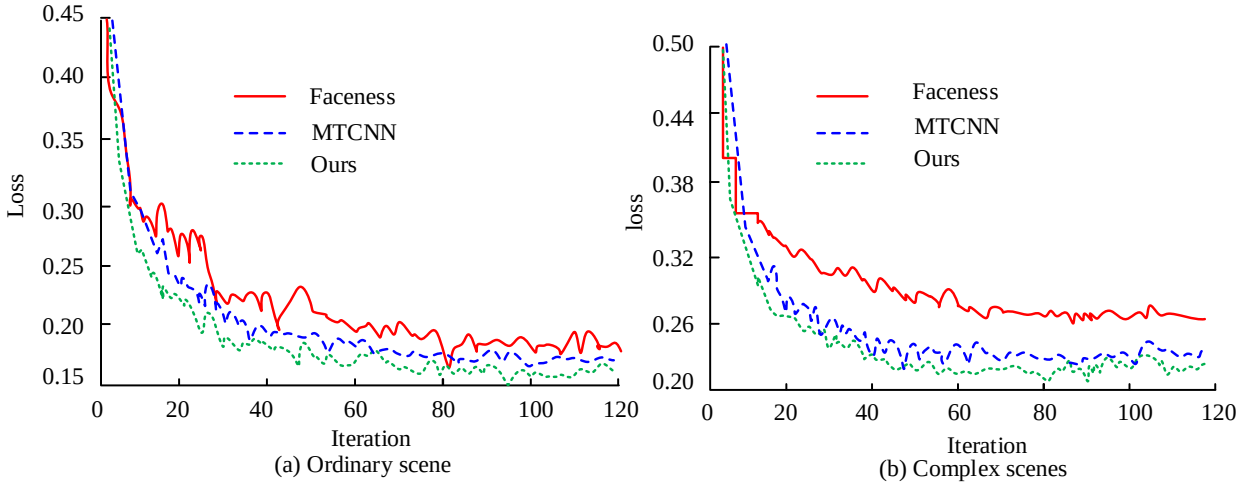
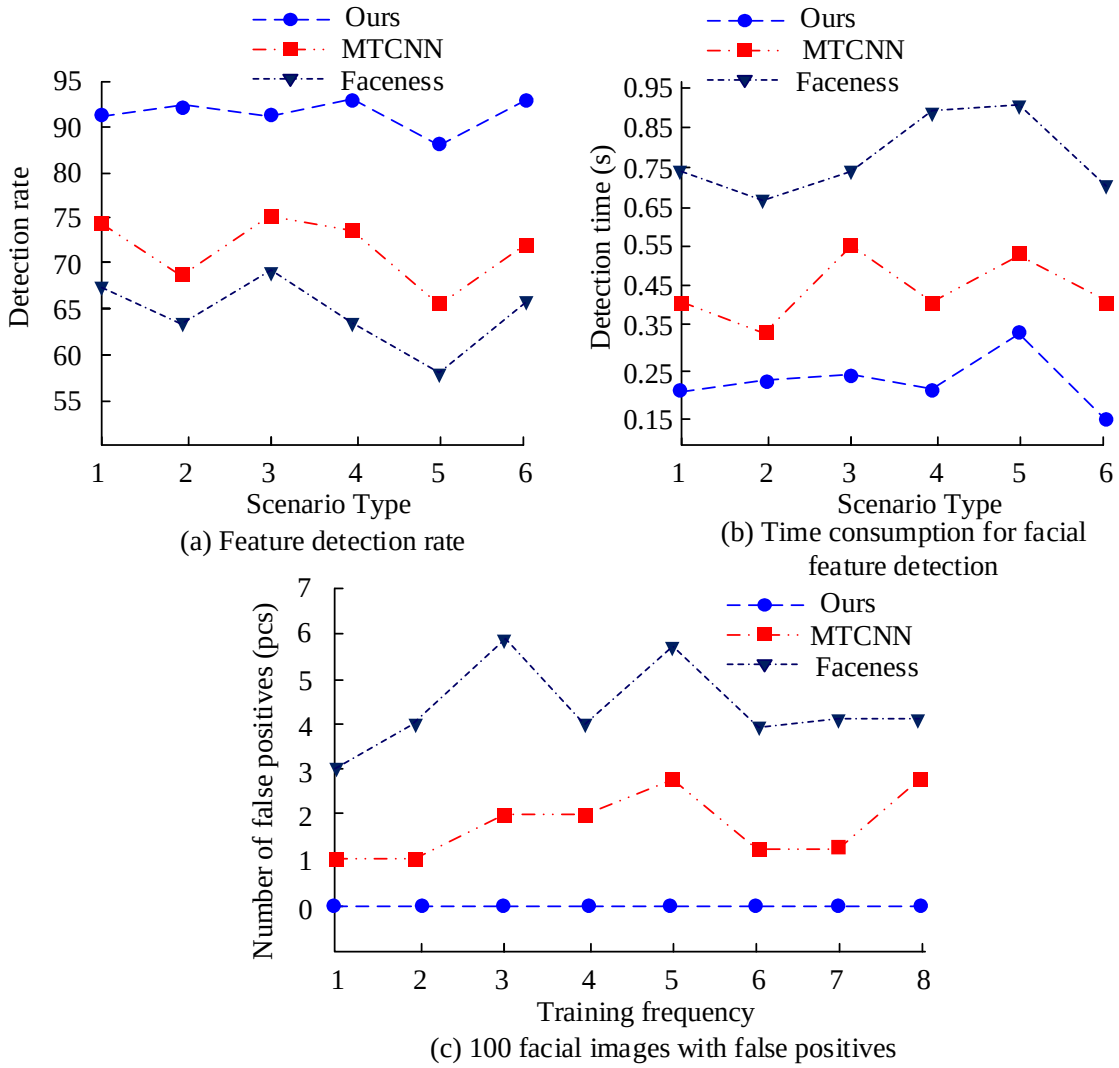


Figure 9: Face detection training loss

Figure 9(a) shows the results of training loss for ordinary scenes. Among them, only the research model has the lowest loss of 0.165 at convergence, while the training loss of MTCNN and Faceness at convergence is 0.184 and 0.205, respectively. In Figure 9(b), which shows the training loss of the complex scene, Faceness has the worst overall training effect and shows obvious fluctuations during the training process. The loss at convergence is 0.286, while MTCNN and the research model perform significantly better, 0.256 and 0.221, respectively. Next, the study uses CelebA data to test the feature detection rate of different face detection techniques and the number of false face detections. Among them, the detection rate indicates the ratio of detected face features to total features, while the number of false detections indicates the number of face judgment errors. The experiment sets up six facial recognition scenarios. Scene 1 is a low-light environment with a light intensity of $\leqslant$50 lux for facial detection and recognition. Scene 2 is a strong backlight scene, where the face is in front of a strong light source (such as sunlight or spotlights) background. Moreover, the facial brightness is $\leqslant$ 100 lux, while the backlight ambient light is $\geqslant$ 10000 lux. Scene 3 is a partially occluded scene, where the face is partially occluded by objects (masks, sunglasses, hands) (covering 30% to 50% of the face), simulating character camouflage or temporary occlusion in movies and TV shows. Scene 4 is a high angle deflection scenario, where the face has significant deflection (side face, pitch) relative to the camera. This exceeds the conventional frontal recognition range. Among them, the horizontal deflection angle is $\geq$ 45 ° or the vertical tilt angle is $\geq$ 30 °. Scene 5 is a motion blurred scene, where facial images are blurred due to rapid movement or camera shake, simulating action scenes or handheld shooting. Scene 6 is a small-sized face scene, and long-distance shooting results in a small proportion of the face in the image and insufficient detail resolution. The specific results are shown in Figure 9.



(a) Feature detection rate

(b) Time consumption for facial feature detection

(c) 100 facial images with false positives

Figure 10: Comprehensive testing of face detection

Figure 10(a) shows the results of feature detection rate. Among the six face scenarios, the research model has the highest detection rate for all types of face features with an average value of 93.54%. MTCNN, which is the next best performer, has an average value of 71.25%. Faceness is only 63.25%. Figure 10(b) shows the time-consuming face feature detection. Overall, Faceness face detection takes the longest time, with a mean value of 0.765s. It is followed by MTCNN, with a mean value of 0.452s. The research model is the shortest, with only 0.253s. Finally, Figure 9(c) shows the final false detection rate result for one hundred faces. The number of false detections represents the number of misidentifications among the recognized faces, measured in individuals (pcs). The training frequency refers to the number of times a model is repeatedly trained on specific interference samples (such as background textures that are prone to false positives) during the training process. The mean value of eight repetitions of the test is 4.37 misdetections per hundred compared to 1.75 for MTCNN. The research model does not suffer from the problem of misdetection and the number of misdetections is zero.

3.2 Face replacement experiment

Next, the face replacement experiments are continued. The study uses a high-resolution face dataset (CelebAMask-High Quality, CelebAMask-HQ) and a specialized video face dataset (Deepfake Model Attribution Dataset, FDM) for the experiments. Among them, CelebAMask-HQ dataset contains 30,000 face data of various types with a mask size of 512×512, which is suitable for face switching experiments. FDM contains 6,450 videos of various types of faces. For testing, the study introduces attribute-preserving generative adversarial network (AP-GAN) and FaceSwap as test benchmarks. Figure 11 first compares the effects of several face replacement approaches using the CelebAMask-HQ dataset.

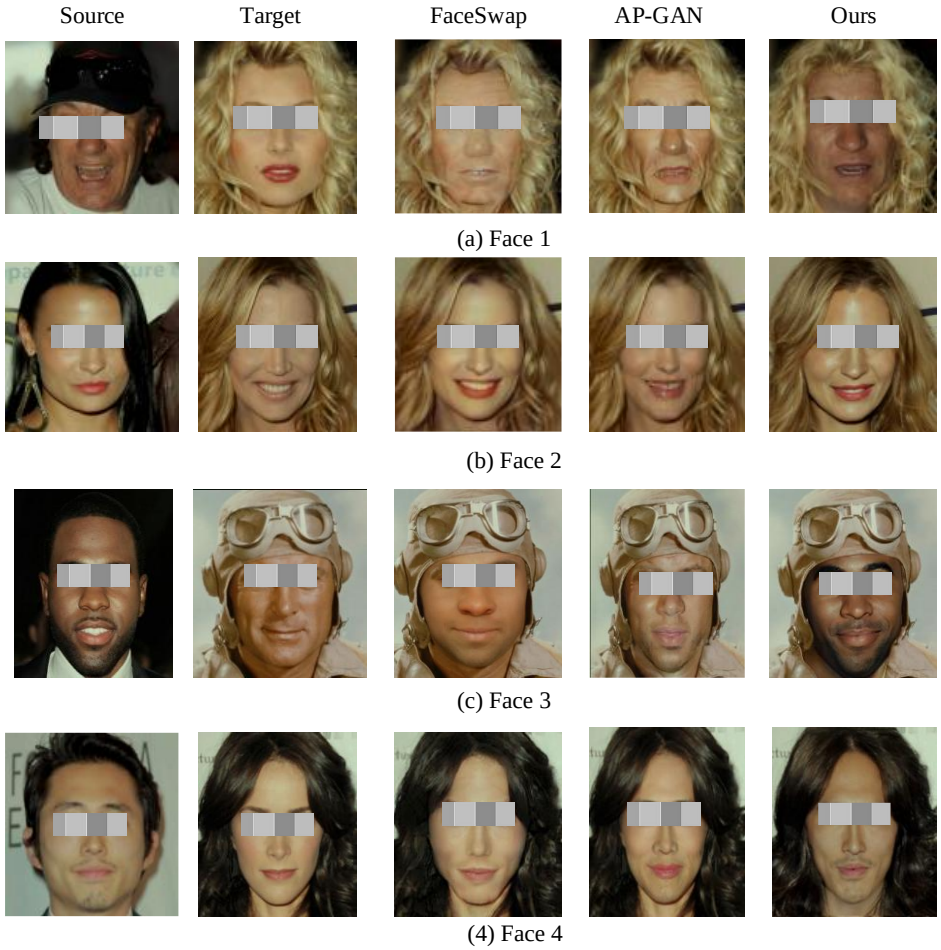
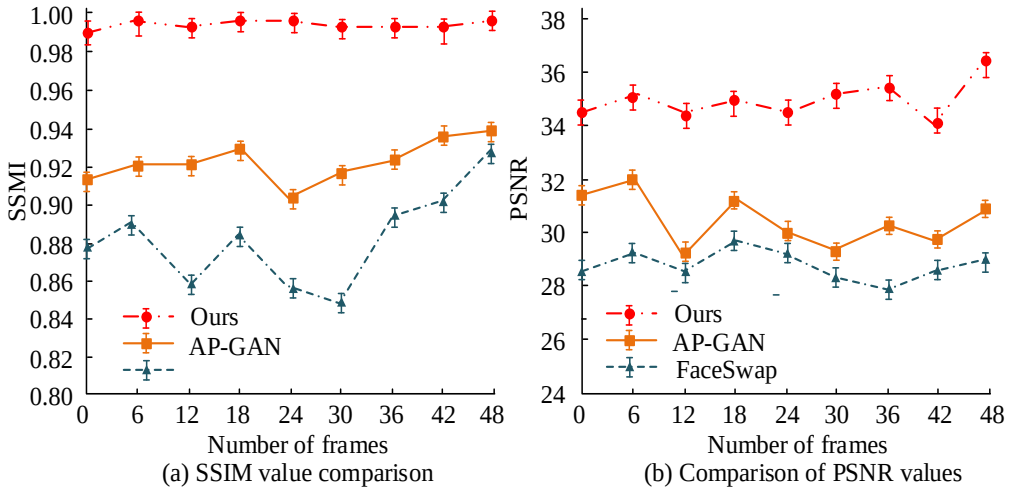


Figure 11: Face replacement effect

In the face image in Figure 11(a), FaceSwap face replacement has problems with skin color differences and mouth shape deviation, while AP-GAN is overall better, but still has problems with skin color deviation as well as jaw overexposure. Only the study that restores the model to the original face contour and skin color performs the best. In Face 2 of Figure 11(b), again only the research model does not show face contour bias as well as color problems. In Face 3 of Figure 11(c), FaceSwap and AP-GAN show the problem of whiteness of face skin color, and FaceSwap face contour is obviously abnormal. The research model, on the other hand, effectively restores the skin color and contour of the original image. In Figure 11(d) of Face 4 the research model has the best color and contour restoration, while all other techniques show contour and color deviations. Next, under the FDM dataset, the study is introduced to introduce structural similarity index (SSMI) and peak signal-to-noise ratio (PSNR) to compare the face replacement effect of different techniques, as shown in Figure 12.



(a) SSIM value comparison (b) Comparison of PSNR values

Figure 12: Comparison between SSMI and PSNR

Figure 12(a) shows the SSIM test results. The research model significantly outperforms the other techniques. In the 0 to 48 frames count scenario, the research model has an average SSMI value of 0.994. In contrast, the performance in AP-GAN face replacement is average, with an average SSMI value of 0.924. The worst performer, FaceSwap, has an average SSMI value of 0.875. Figure 12(b) shows the results of the PSNR comparison. Similarly in the 0-frame to 48-frame face replacement scene, only the research model has a high PSNR value, indicating that it reconstructs faces with better quality. The average PSNR values of the research model, AP-GAN, and FaceSwap are 35.65, 31.05, and 28.84, respectively. The FDM face data is selected to compare the effect of different techniques face replacement. The state error measures the degree of matching of facial dynamic attributes, including expressions, poses, and motion coherence. For video face swapping to look natural, the replaced face must be synchronized with the target video's dynamic changes (e.g., blinking or turning the head). Otherwise, there will be uncoordinated "shaking" or "lag." attribute errors reflect the ability to preserve static identity features, including inherent attributes such as skin color, facial features, facial contours, etc. The higher the error, the lower the similarity between the replaced face and the source identity (such as skin whitening, contour deformation). The details are shown in Table 4.

Table 4: Comprehensive testing of face replacement using different technologies

Face replacement number	FaceSwap			AP-GAN			Ours		
	Structural similarity	State error	Attribute error	Structural similarity	State error	Attribute error	Structural similarity	State error	Attribute error
1	0.726	3.548	0.458	0.851	3.868	0.322	0.982	2.541	0.124
2	0.756	4.055	0.425	0.826	2.956	0.357	0.986	2.057	0.114
3	0.743	3.853	0.405	0.816	3.457	0.353	0.991	2.043	0.113
4	0.735	3.754	0.425	0.788	3.054	0.342	0.986	2.044	0.125
5	0.684	3.686	0.432	0.816	2.985	0.325	0.991	2.123	0.135
6	0.765	4.285	0.432	0.823	2.973	0.325	0.998	2.255	0.126
7	0.726	3.785	0.456	0.838	3.054	0.315	0.983	2.034	0.122
8	0.715	4.066	0.428	0.793	3.522	0.334	0.986	2.015	0.155
9	0.706	3.856	0.416	0.834	2.893	0.332	0.993	2.132	0.153
10	0.698	3.852	0.405	0.835	3.085	0.325	0.992	2.053	0.164
Average value	0.725	3.874	0.428	0.822	3.185	0.333	0.989	2.130	0.133

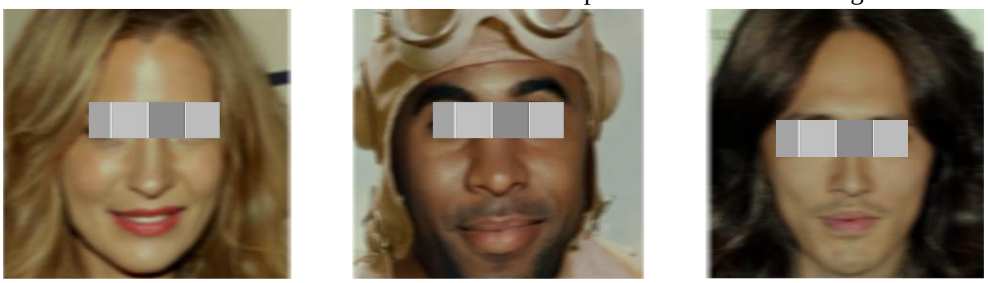
Table 4 shows the comprehensive test results of different techniques face replacement, which are evaluated in terms of SSMI, state error, and attribute error. The results of the 10 sets of face tests show that the research model has the best overall performance. For example, in the SSMI test, the average value of the research model is 0.989, while FaceSwap is 00.725 and AP-GAN is 0.882. In the comparison of state error, the average value of the research model is 2.130, which is better than FaceSwap and AP-GAN's 3.874 and 3.185. In conclusion, the research technique has the best overall performance in face switching. In addition, the study compares the anti jitter effects of research models, using frame continuity measurement (FCM) and motion consistency error (MCE) as core indicators. FCM evaluates fluency by calculating the variance of feature point displacement between adjacent frames. The higher the value, the more stable it is. MCE measures the consistency of facial keypoint trajectories. The lower the value, the more consistent the trajectories are. The test results are shown in Table 5.

Table 5: Vibration test

Test indicators	No shaking module	There is a shaking module	*p*-value	Effect size (Cohen's d)
FFCM	0.872±0.032	0.956±0.019	<0.001**	1.24
MCE	3.241±0.215	2.115±0.183	0.003**	0.91
PSNR(dB)	33.72±1.05	35.65±0.87	0.008**	0.78
SSIM	0.931±0.021	0.989±0.012	<0.001**	1.37

According to Table 5, the debounce module in FCM optimizes interframe transitions through optical flow analysis and motion compensation. This results in a significant improvement in FCM, raising the score from 0.872 to 0.956. The comparison between the data is statistically significant, $p<0.001$. Moreover, its standard deviation decreased from 0.032 to 0.019, indicating that the module effectively reduced inter frame fluctuations. The standard deviation of MCE decreases from 0.215 to 0.183 in the MCE test, a 34.7% decrease, indicating that the module effectively suppresses the random shaking component of facial motion. In addition, the image generation quality test revealed significant

improvements in both PSNR and SSIM values, with a statistically significant difference of $p<0.001$ between the before and after results. This module controls the inter frame coefficient of variation within an industrial threshold (CV<2%), providing key technical support for real-time face replacement at the film and television level. The comparison of the debounce effect before and after face replacement is shown in Figure 13.



(a) Not shaking yet

(b) After shaking

Figure 13: Comparison of face replacement and debounce effects

Figure 13 (a) shows the face replacement effect without debounce, while Figure 13 (b) shows the image result after face replacement and debounce. Clearly, the image without debounce has obvious issues with boundary ghosting, contour blurring, and hair texture blurring. After deblurring, the image has prominent edge contours, clear color blocks, and a significantly higher level of brightness and darkness. The research continues to conduct ablation experiments on the debounce module, as shown in Table 6.

Table 6: Experiment on the ablation of the shaking module

Evaluation indicators	No processing	RSC module only	RSC+Smooth Filter
MCE	4.52±0.31	3.15±0.24	2.13±0.18
FCM	0.81±0.04	0.91±0.02	0.96±0.01
PSNR(dB)	31.85±1.12	34.20±0.95	35.63±0.87
SSIM	0.902±0.025	0.962±0.018	0.988±0.012
Inter frame difference coefficient (%)	8.7±1.2	4.3±0.8	1.4±0.3

Table 6: shows the results of the ablation experiment on the debounce module. According to the test results, no task module processing is performed, and the MCE, FCM, PSNR, and SSIM values are all low, with a high frame rate difference coefficient of 8.7±1.2%. After adding the RSC module, the MCE value decreases significantly, from 4.52 to 3.15. Meanwhile, the FCM value increases, from 0.81 to 0.91. This change is mainly due to the enhanced stability of optical flow tracking. In addition, the PSNR and SSIM values for image generation quality increases significantly. In contrast, the MCE and FCM values for RSC+Smooth Filter are 2.13±0.18 and 0.96±0.01, respectively, indicating a significant decrease in image generation quality. This result also indicates that adding RSC module and Smooth Filter to the model can significantly improve the face replacement effect. Finally, the research introduces a comprehensive comparison of the computational efficiency and processing effectiveness of different models under resource constraints or abundance using literature [7] and UGAN Lite technology. Among them, the resource rich scenario is NVIDIA RTX3070 GPU (12GB video memory), batch size=16. The running memory is 64GB. The resource constrained scenario is NVIDIA RTX3070 GTX1060 (6GB of video memory), with a running memory of 16GB. The comparison of computing efficiency is shown in Table 7.

Table 7: Comparison of calculation efficiency and processing effect of different models

Model	Environment	Reasoning time (ms)	Memory footprint (MB)	FLOPs (G)	Frame rate (fps)
Ours	Rich resources	21.5±1.2	345±25	19.8	43.5
	Resource-constrained	128.3±8.5	245±30	15.8	15.4
AP-GAN	Rich resources	35.6±1.0	354±20	12.4	35.8
	Resource-constrained	152.4±5.1	246±25	7.4	11.5
FaceSwap-GAN	Rich resources	48.5±0.5	352±20	13.8	38.5
	Resource-constrained	254.5±2.3	254±34	9.4	10.8
Reference [7]	Rich resources	48.2±1.8	378±35	13.2	38.5
	Resource-constrained	275.5±9.2	264±28	11.5	12.4
UGAN-Lite	Rich resources	31.4±0.8	367±18	14.8	40.5
	Resource-constrained	156.4±1.5	258±84	11.8	14.4

As shown in Table 7, there are significant differences in the computational efficiency of the model between resource-rich and resource-limited scenarios, including the inference time, throughput (FLOPs), and processing frame rate. Under abundant resources, different models have better computational efficiency and higher processing frame rates. Among them, the research model performs the best, with the shortest inference time of 21.5 ± 1.2ms in the resource rich state, which is better than UGAN Lite's 31.4±0.8ms. Additionally, the research model shows the best performance in the PLOPs ratio under abundant and resource-limited conditions, with throughputs of 19.8G and 15.8G, respectively. In terms of processing frame rates, the research model performs significantly better than similar technologies under both resource-limited and resource-rich conditions, with processing frame rates of 15.4 fps and 43.5 fps, respectively. For example, the processing frame rates of the technology proposed in reference [7] under resource rich and resource limited conditions are 38.5 and 12.4, respectively. In summary, the research model has computational efficiency in different resource scenarios, and the task processing effect is better.

4 Discussion

In recent years, face replacement technology has been widely applied in fields such as film and video production, advertising design, etc. This is due to the rapid development of AI and DL technology. However, this technology still faces many challenges, such as insufficient accuracy and high resource consumption. A study proposed an automated face replacement technique based on improved MTCNN and GAN to address these issues.

In the face detection experiment, the improved MTCNN algorithm performed excellently. In ordinary scenarios, the accuracy of the research model reached 98.35%, significantly higher than the 92.31% of traditional MTCNN and 82.65% of Faceness. In complex scenarios, the accuracy of the research model was 93.25%, which was better than MTCNN's 86.25% and Faceness's 81.25%. This indicated that the improved MTCNN algorithm improved face detection accuracy and efficiency by introducing median filtering and depth separation convolution. In terms of training loss, the research model's convergence loss was 0.165 and 0.221 in ordinary and complex scenarios, respectively. These values were both lower than those of other techniques. In the face replacement experiment, the research model also performed well. In terms of preserving color and facial contour, the research model was more effective at restoring the original image's facial features, avoiding issues like skin color deviation and contour irregularities. In the SSMI test, the average value of the research model was 0.994, significantly higher than AP-GAN's 0.924 and FaceSwap's 0.875. In the PSNR test, the average value of the research model was 35.65, which was better than AP-GAN's 31.05 and FaceSwap's 28.84. This indicated that the face replacement model based on adversarial generative networks used a SAM and a digital debounce method. These features significantly improved the accuracy of face replacement and the smoothness of video processing.

Further analysis revealed that the application of self attention mechanism in face replacement was particularly crucial. It improved the model's ability to focus on important features by assigning weights to them, thereby making the replacement more natural. The digital debounce module effectively solved the jitter problem in video replacement through corner detection, optical flow analysis, and motion estimation, making video frames smoother. In terms of computational trade-offs, the research model performed well in both resource-rich and resource-limited scenarios. It had better inference time, throughput, and processing frame rate than similar technologies.

However, it should be noted that while the use of deepfake technology in film and television is innovative, its misuse poses social risks and raises ethical concerns that cannot be ignored. The abuse of Deepfake

technology may lead to the spread of false information, infringement of personal privacy and portrait rights. For example, unauthorized facial replacement may be used to create fake videos for illegal activities such as fraud and defamation. In addition, the rapid development of Deepfake technology has made forged videos increasingly realistic, making them difficult to distinguish as fake with the naked eye. This not only poses a threat to personal safety, but also impacts the social trust system. Therefore, when using relevant technologies, it is necessary to comply with local laws and regulations. Service providers of deep synthesis technology must fulfill their prompting and supervisory obligations to ensure the technology's safe use, as well as abide by ethical and moral principles.

In summary, the proposed facial replacement technology has demonstrated excellent accuracy, efficiency, and naturalness. This technology provides strong technical support for fields such as film and video production.

5 Conclusion

In recent years, with the continuous development of AI and DL, face replacement technology has become a focus of attention. Therefore, the research proposed an automated face replacement technology. The technology adopted an improved MTCNN algorithm for face detection, which improved the accuracy and efficiency of face detection by introducing median filtering and depth separation convolution. During the face replacement stage, a UGAN-based model was constructed that combined a SAM and a digital debounce method. This improved the accuracy of the face replacement and the smoothness of the video. In the face detection experiment, the improved MTCNN algorithm performed well. In ordinary scenarios, the accuracy of the research model reached 98.35%, significantly higher than the 92.31% of traditional MTCNN and 82.65% of Faceness. In complex scenes, the accuracy of the research model was 93.25%, which was better than MTCNN's 86.25% and Faceness's 81.25%. In the face replacement experiment, the research model also performed well. In terms of preserving color and facial contour structure, the research model could better restore the facial features of the original image, avoiding problems such as skin color deviation and contour anomalies. In the SSMI test, the average value of the research model was 0.994, significantly higher than AP-GAN's 0.924 and FaceSwap's 0.875. In the PSNR test, the average value of the research model was 35.65, which was better than AP-GAN's 31.05 and FaceSwap's 28.84. In summary, it can be seen that the technology proposed by the research has good application effects in video face switching. However, the research has shortcomings as well. For example, UGAN-based face replacement technology cannot process large amounts of offline video data. In addition, GANs also face the problem of hallucinations that occur in occlusions or rare poses during training. In the future, it will be necessary to develop an offline processing system that can adapt to different video face replacement requirements in various scenes, while improving technical recognition under occlusion and enhancing complex backgrounds.

References

[1] Zeng H, Zhang W, Fan C, Lv T, Wang S. FlowFace: semantic flow-guided shape-aware face-swapping[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(3): 3367-3375. https://doi.org/10.1609/aaai.v37i3.25444

[2] Rehaan M, Kaur N, Kingra S. Face manipulated deepfake generation and recognition approaches: A survey. Smart Science, 2024, 12(1): 53-73. https://doi.org/10.1080/23080477.2023.2268380

[3] Kalembo Vikalwe Shakrani, Ngonidzashe Mathew Kanyangarara, Prince Tinashe Parowa, Vibhor Gupta, Rajendra Kumar. A Deep Learning Model for Face Recognition in Presence of Mask. Acta Informatica Malaysia. 2022; 6(2): 43-46. https://doi.org/10.26480/aim.02.2022.43.46

[4] Rao I S S, Kumar J S, Vamsi T M N, Kumar TR. IMPROVING REALISM IN face-swapping USING DEEP LEARNING AND K-MEANS CLUSTERING. Proceedings on Engineering, 2024, 6(4): 1751-1756. https://doi.org/10.24874/PES.SI.24.03.010

[5] Omar K, Sakr R H, Alrahmawy M F. An ensemble of CNNs with self-attention mechanism for DeepFake video detection. Neural Computing and Applications, 2024, 36(6): 2749-2765. https://doi.org/10.1007/s00521-023-09196-3

[6] AbdElminaam D S, Sherif N, Ayman Z, et al. DeepFakeDG: A deep learning approach for deep fake detection and generation. Journal of Computing and Communication, 2023, 2(2): 31-37. https://doi.org/10.21608/jocc.2023.307056

[7] Tsai C S, Wu H C, Chen W T, Ying JJC. ATFS: A deep learning framework for angle transformation and face-swapping of face de-identification. Multimedia Tools and Applications, 2024, 83(12): 36797-36822. https://doi.org/10.1007/s11042-023-16123-0

[8] Melnik A, Miasayedzenkau M, Makaravets D. Face generation and editing with stylegan: A survey. IEEE Transactions on pattern analysis and machine intelligence, 2024, 46(5): 3557-3576. https://doi.org/10.1109/TPAMI.2024.3350004

[9] Liao X, Wang Y, Wang T, Hu J, Wu X. FAMM: Facial muscle motions for detecting compressed deepfake videos over social networks. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(12): 7236-7251. https://doi.org/10.1109/TCSVT.2023.3278310

[10] Akhtar Z, Pendyala T L, Athmakuri V S. Video and audio deepfake datasets and open issues in deepfake technology: being ahead of the curve. Forensic Sciences, 2024, 4(3): 289-377. https://doi.org/10.3390/forensicsci4030021

[11] Zhao H, Zhou W, Chen D, Zhang W, Guo Y. Audio-Visual Contrastive Pre-train for Face

Forgery Detection. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 21(2): 1-16. https://doi.org/10.1145/3651311

[12] Yue P, Chen B, Fu Z. Local region frequency guided dynamic inconsistency network for deepfake video detection. Big Data Mining and Analytics, 2024, 7(3): 889-904. https://doi.org/10.26599/BDMA.2024.9020030

[13] Wöhler L, Ikehata S, Aizawa K. Investigating the Perception of Facial Anonymization Techniques in 360° Videos. ACM Transactions on Applied Perception, 2024, 21(4): 1-17. https://doi.org/10.1145/3695254

[14] Lai Y. Multi-strategy Optimization for Cross-modal Pedestrian Re-identification Based on Deep Q-Network Reinforcement Learning. Informatica, 2025, 49(11). https://doi.org/10.31449/inf.v49i11.7247

[15] Yang G, Xu K, Fang X, Zhang J. Video face forgery detection via facial motion-assisted capturing dense optical flow truncation. The Visual Computer, 2023, 39(11): 5589-5608. https://doi.org/10.1007/s00371-022-02683-z

[16] Pang G, Zhang B, Teng Z, Qi Z. MRE-Net: Multi-rate excitation network for deepfake video detection. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 3663-3676. https://doi.org/10.1109/TCSVT.2023.3239607

[17] Deng L, Wang J, Liu Z. Cascaded network based on EfficientNet and transformer for deepfake video detection. Neural Processing Letters, 2023, 55(6): 7057-7076. https://doi.org/10.1007/s11063-023-11249-6

[18] Liu T. Secure Face Recognition Using Fully Homomorphic Encryption and Convolutional Neural Networks. Informatica, 2024, 48(18). https://doi.org/10.31449/inf.v48i18.6396

[19] Lu T, Bao Y, Li L. Deepfake Video Detection Based on Improved CapsNet and Temporal-Spatial Features. Computers, Materials and Continua, 2023, 75(1): 715-740. https://doi.org/10.32604/cmc.2023.034963

[20] Pang G, Zhang B, Teng Z. MRE-Net: Multi-rate excitation network for deepfake video detection. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 3663-3676. https://doi.org/10.1109/TCSVT.2023.3239607